The Connecticut Academic Performance Test: Technical Report

Prepared by
Irene Hendrawan & Arianto Wibowo

October 2008

Table of Contents

List of Charts

List of Tables

Part 1: Introduction

1.1. General description of CAPT

The Connecticut Academic Performance Test (CAPT) was designed to measure student performance in high school. Students are tested in the areas of Mathematics, Science, Reading, and Writing.

The CAPT has measured achievement of Connecticut students since 1994, when it was first administered. The second generation of CAPT was introduced in 2001. The content structure of the first generation CAPT was used as the baseline in developing the second generation. For the most part, the educational outcomes tested in the first generation were carried over to the second generation. Changes were made in light of new trends in instruction, educational assessment, and the lessons learned over the years of the first generation. The third generation of CAPT was introduced in the spring of 2007. The spring 2008 administration was the second operational (OP) administration of CAPT3

1.2. 2008 CAPT Test Design

The spring 2008 administration comprises the following content areas:

1. Mathematics
   Mathematics (MA) has thirty-two operational -- twenty-four grid-in (GR) response items and eight open-ended (OE) items scored on 0-3 scale.
2. Science
   Science (SC) has sixty-five OP items -- sixty multiple choice (MC) items and five OE items scored on 0-3 scale.
3. Reading
   Reading (RD) consists of two subtests:
   - Reading for Information
     Reading for Information (RI) has eighteen OP items -- twelve MC items and six OE items scored on 0-2 scale.
   - Response to Literature
     Response to Literature (RL) consists of an extended response (EX) item with a 2-12 score scale (sum of two rater scores on a 1-6 scale).
4. Writing
   Writing (WR) consists of three subtests:
   - Editing & Revising
     Editing & Revising (ER) has eighteen MC items.
   - Interdisciplinary Writing 1 & Interdisciplinary Writing 2
     Interdisciplinary Writing 1 (IW1) & Interdisciplinary Writing 2 (IW2) have an EX item with a 2-12 score scale (sum of two rater scores on a 1-6 scale).

Table 1: 2008 CAPT Operational Test Design

Content Area	Subject	Number of Items				Total Items	Raw Score
		MC	GR	OE	EX		
Mathematics	Mathematics		24	8		32	0 - 48
Science	Science	60		5		65	0 - 75
Reading	Reading for Information	12		6		18	0 - 24
	Response to Literature				1	1	2 - 12
Writing	Editing & Revising	18				18	0 - 18
	Interdisciplinary Writing 1				1	1	2 - 12
	Interdisciplinary Writing 2				1	1	2 - 12

1.3. 2008 CAPT Test Forms

In the 2008 administration, two main forms were available for administration: Form HS15, which is the live form taken by most of the students, and Form HS0, which was available for breach situations. Moreover, Form HS0 will be used as a breach form in subsequent years of the third generation. Although the two forms were pre-equated during test assembly, there was still a need to carry out a post equating procedure after the test administration in order to ensure the comparability of the two forms.

For Writing, post equating was accomplished by using common test sessions or subtests that connected the two forms. Two stratified samples of 2,000 students were selected to take two hybrid forms that consisted of part of Form HS15 (Live) and part of Form HS0 (Breach). Scores on the hybrid forms constituted the live scores for the members of the two samples. Table 2 shows how these hybrid forms were built.

Table 2: 2008 CAPT Test Forms

Form	Writing	N
HS15 - Live	IW1 HS15	Most of the State (approximately 40,000)
	IW2 HS15	
	ER HS15	
Hybrid 1	IW1 HS15	2,000
	IW2 HS0	
	ER HS15	
Hybrid 2	IW1 HS0	2,000
	IW2 HS15	
	ER HS15	
HS0 - Breach	IW1 HS0	Hybrid or Breach cases
	IW2 HS0	
	ER HS0	

The stratification of the samples for the hybrid forms was based on scale score distribution. CSDE's rationale for stratifying the test forms based on scale scores from the previous year was that this procedure would more likely yield groups of test takers who were representative with respect to the distribution of skills and achievement across the entire state. In other words, instead of sampling based on conventional demographic variables to achieve representation of test-taker characteristics, CSDE chose to sample on test-taker achievement. MI selects a stratified sample of schools, based on the scale score distribution to which each belongs.

Any student who breaches a test session or subtest (HS15 or HS0) was given the corresponding test session or subtest (HS15 or HS0).

Part 2: Test Development

The process by which each form of the CAPT is developed is extensive, spanning a five- or six-year period and many stages. The development process is led and overseen by staff members in the Bureau of Student Assessment at the Connecticut State Department of Education (CSDE), but it also involves many other people who represent a wide variety of perspectives and areas of expertise. CSDE curriculum specialists and content experts play a critical role and work closely with the assessment staff throughout the process. In addition, a major testing company and other organizations and individuals with experience in educational assessment are involved at appropriate points in the development process.

Advisory committees of Connecticut educators are particularly important throughout the development of the CAPT. Content Advisory and Fairness Committees review each item to ensure the match between the content objectives and the items, and to ensure meaningful interpretability of test results. The Content Advisory Committees included content experts, regular and special education teachers, Connecticut State Department of Education curriculum, and assessment content specialists. A separate advisory committee is established for each part of the CAPT: Mathematics, Science, Reading, and Writing. These advisory committee members are selected on the basis of their knowledge in educational content and processes. In addition, the Fairness Committee is responsible for determining whether items are appropriate and fair to all examinees. Educators are carefully selected for the advisory committees to be representative of school districts throughout Connecticut.

The test development process for CAPT3 began with content specialists and testing experts writing test specifications with the help of the CAPT content advisory committees. The starting point for this process was looking at the specifications and structure of the first generation CAPT, and examining what has been working and what needed improvement. The new curriculum frameworks adopted by the State of Connecticut were also used as a guide. Test items for the CAPT3 were carefully developed in accordance with the established test specifications and test blueprint. These items were carefully matched to the content standards in the Connecticut Curriculum Frameworks for Mathematics, Science, Reading, and Writing. Items that did not pass the scrutiny of either Content Advisory or Fairness Committees were eliminated from the pool of pilot items.

After committee reviews, field test forms were created and piloted on a representative sample, stratified by scale score distribution, consisting of approximately 2000 students per form. Pilot statistics such as the mean, point biserial, and Rasch difficulty were generated and reviewed by CSDE assessment content staff and psychometricians. In addition, for hand-scored constructed response items, the contractor staff provided qualitative summaries about whether students appeared to have sufficient contextual knowledge to be able to fully respond to the item. Flawed items were removed from the item pool, including those showing test item bias or inappropriate levels of difficulty. Based on the CAPT3 Blueprints, Mathematics, Science, Reading, and Writing test forms of equivalent difficulty were simultaneously constructed from the pool of items that met all the review criteria. Every effort was made to ensure that strand level difficulties were comparable and that the items reflected the appropriate range of content within the strands across the generation.

Part 3: Item Level Statistics

Tables 3 and Appendix A present item analysis (item quality) data for Mathematics, Science, Reading and Writing. The following information is presented in each item analysis:

Classical and IRT difficulties: Item difficulty is fundamentally a ratio of the proportion of examinees who answered the item correctly to those who did not answer the item correctly. Thus, an easy item has a high p-value and a difficult item has a low p-value. If an item has a very high p-value it may be so easy that it does not provide much information about what most examinees know or can do, while an item with a very low p-value may be so difficult that it is beyond the range of what most students know or can do. Therefore, items with very high or very low p-values may be rejected, unless content relevance overrides that concern.

Item Discriminations: The point biserial correlation or item-total correlations measure the strength of the relationship between the particular item score and the total score. Thus, item discrimination reflects how well a particular item differentiates between high and low total test performers. When the correlation is high, examinees that do well on the item also tend to do well on the entire test and correspondingly, examinees that do not do well on the item also tend not to do well on the total test.

Distractor Frequencies: The proportion of students who answered each option (A-D, 0-3, and 2-12) are presented for the multiple-choice items, open-ended and extended response, respectively. The percent of students at each score point is presented for extended response (2-12).

Table 3: Summary of Item Analysis Form HS15

Subject	Rasch		P-value		Point Biserial	
	Mean	Std	Mean	Std	Mean	Std
Mathematics	0.1119	0.7018	0.71	0.43	0.56	0.09
Reading for Information	-0.2073	1.3210	0.71	0.16	0.40	0.07
Response to Literature	0.4173	.	7.08	.	0.62	.
Editing and Revising	0.1614	1.2347	0.68	0.19	0.28	0.07
Interdisciplinary Writing	1.3556	0.1285	7.36	0.21	0.70	0.01
Science	0.0045	0.7106	0.69	0.31	0.40	0.10

Part 4: Scaling and Equating

4.1 2008 CAPT Linking Items

The 2008 CAPT Mathematics, Science, Reading for Information, and Editing & Revising tests were equated with the 2007 CAPT (HS14) subtests by embedding linking items which were carried over from the live 2007 (HS14) administration. Linking items were counted toward students' scores.

The Live form of the 2008 CAPT (HS15) included:

- Mathematics – twelve linking grid items were embedded.
- Science – fifteen linking MC items were embedded.
- Editing & Revising – one passage with six linking MC items were embedded.
- Reading for Information – one passage with four linking MC items and two linking OE items.

Table 4 indicates the linking items used as well as their positions on the 2007 and 2008 tests.

Table 4: 2008 Embedded Linking Items

Note: Items dropped after anchor evaluation are shaded (see section 4.2.)

Content Area	Form HS14 A Item Position	Form HS15 Item Position	Item Type	Rasch Form HS14
Mathematics	5	5	GR	-0.3799
	6	6	GR	-0.3157
	7	7	GR	0.1066
	13	12	GR	0.8691
	15	14	GR	-0.2066
	18	16	GR	-0.5274
	24	23	GR	0.3959
	26	25	GR	-0.6782
	27	26	GR	0.4089
	32	29	GR	0.9282
	33	30	GR	-0.4329
	36	32	GR	-0.2411
Science	6	6	MC	-0.5395
	7	7	MC	0.5883
	9	8	MC	-0.2542
	14	14	MC	0.5723
	15	15	MC	0.6178
	17	16	MC	0.7519
	53	51	MC	-0.2104
	55	52	MC	0.7209
	54	53	MC	-0.5523
	63	60	MC	0.5610
	64	61	MC	-0.2141
	66	62	MC	0.1998
	67	63	MC	-0.0207
	68	64	MC	-0.5744
	69	65	MC	-0.0877

Content Area	Form HS14 A Item Position	Form HS15 Item Position	Item Type	Rasch Form HS14
Reading	7	7	MC	-0.7558
	8	8	MC	-0.0018
	9	9	MC	-1.2833
	10	10	MC	-0.1870
	11	11	OE	0.7861
	12	12	OE	1.1143
Writing	13	1	MC	0.7168
	14	2	MC	-0.9479
	15	3	MC	1.7167
	16	4	MC	-0.0903
	17	5	MC	0.9715
	18	6	MC	-1.0287

4.2. Calibration Process

The CAPT 2008 tests were scaled and equated using the Rasch model. The WINSTEPS software was used to estimate the latent trait difficulty of each item on the test. WINSTEPS, written by Linacre (Mesa Press, 2005) was used to complete Rasch analyses. WINSTEPS is a WINDOWS-based program that is widely used for similar high stakes tests. WINSTEPS (the Rasch model), allows for the estimation of item difficulty for multiple-choice, open-ended, and extended response items on a single scale. Using these item difficulties, the model is able to estimate the ability (theta) of each student corresponding to each student's raw score.

All scaling and equating analyses were undertaken by three independent groups: Measurement Incorporated (MI), the contractor, the Connecticut State Department of Education (CSDE), and H. Jane Rogers and H. Swaminathan from the University of Connecticut (UCONN). Results were compared and cross-checked to the fourth decimal point to ensure accuracy.

The purpose of equating was to place the difficulty estimates of the items on the same scale as HS14 (CAPT 2007 Live). The equating was accomplished in the following steps:

1. For every content area, concurrently calibrate the 2008 OP (see Charts 1-4 for sample calibration data matrix). This step is a free run calibration. For RL, IW1, and IW2, 2 is subtracted from each score so that scores are on a scale from 0 to 10.

Chart 1: Calibration Design for 2008 Mathematics

Form HS15	HS15_MA1	HS15_MA2

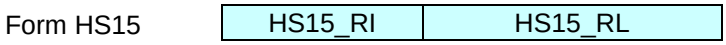
Note:
HS15_MA1 = Form HS15 Math Session 1
HS15_MA2 = Form HS15 Math Session 2

Chart 2: Calibration Design for 2008 Science

Form HS15	HS15_SC1	HS15_SC2

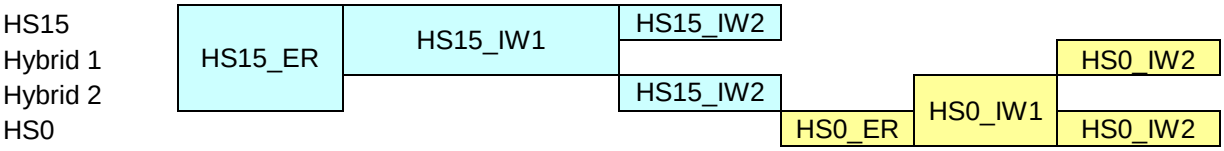
Note:
HS15_SC1 = Form HS15 Science Session 1
HS15_SC2 = Form HS15 Science Session 2

Chart 3: Calibration Design for 2008 Reading



Note:
HS15_RI = Form HS15 Reading for Information
HS15_RL = Form HS15 Response to Literature

Chart 4: Calibration Design for 2008 Writing



Note:
HS15_ER = Form HS15 Editing & Revising
HS15_IW1 = Form HS15 Interdisciplinary Writing 1
HS15_IW2 = Form HS15 Interdisciplinary Writing 2
HS0_ER = Form HS0 Editing & Revising
HS0_IW1 = Form HS0 Interdisciplinary Writing 1
HS0_IW2 = Form HS0 Interdisciplinary Writing 2

2. Select the items linking HS15 (2008 live test) and HS14 (2007 live test). Do anchor evaluation using .3 rule between the estimates of difficulties from Step 1 and HS14 values (see Table 4 for the Rasch values of linking items). This is an iterative process in which each item, starting with the one with the greatest absolute value difference, is removed until all items fulfill the criterion for inclusion. Using the remaining items the difference between the scale means from HS14 and Step 1 yields the equating constant. Table 5 shows the equating constants.

Table 5: 2008 CAPT Equating Constants

Content Area	Equating Constant
Mathematics	0.1119
Reading	-0.1744
Science	0.0045
Writing	0.29995

3. Using the item output files from step 1 and anchoring these b-values, perform another run for each combination of forms, i.e., employ only those items from a given form in order to obtain theta values for each group of students administered a particular form. For Reading and Writing, the appropriate weights were included (see Table 6).

Table 6: Summary of Weighting for Reading and Writing

Content/Subject	Unweighted Scale	% of Total Scale	Score Weight	Compute Formula	Weighted Scale
Reading for Information	0 - 24	50%	1.0		0 - 24
Response to Literature	2 - 12	50%	2.4	(RL - 2)*2.4	0 - 24
Total Reading	2 - 36				0 - 48

Content/Subject	Unweighted Scale	% of Total Scale	Score Weight	Compute Formula	Weighted Scale
Editing & Revising	0 - 18	30%	1.0		0 - 18
Interdisciplinary Writing 1	2 - 12	35%	2.1	(IW1 - 2)*2.1	0 - 21
Interdisciplinary Writing 2	2 - 12	35%	2.1	(IW2 - 2)*2.1	0 - 21
Total Writing	4 - 42				0 - 60

4. Compute scale score (SS) and scale score standard error (SSE) for each forms:

$$SS = \left(\frac{T + EQ - T_{mean}}{T_{SD}} \right) * 45 + 250 \text{ and } SSE = \frac{T_{err}}{T_{SD}} * 45$$

where

T and T_{err} are the ability score and the standard error of the ability from the score file in Step 3 (for Reading and Writing) and Step 1 (for Mathematics and Science).

EQ is the difference between the mean of difficulty estimates of the linking items on HS14 and mean of difficulty estimates of the linking items on HS15, called the equating constant. This value was obtained in Step 2.

T_{mean} and T_{SD} are the scaling coefficients from base year of CAPT2 (see Table 7).

Table 7: Scaling Coefficients from Base Year (CAPT2)

Content Area	T_mean	T_SD
Mathematics	-0.2317	1.6051
Science	0.4077	0.9254
Reading	0.4843	1.2278
Writing	1.0931	1.1187

The minimum SS is set to 100 and the maximum SS is set to 400. Any SS less than 100 was reset to 100 and any SS greater than 400 was reset to 400.

Appendix B contains the results of raw scores, theta, and scale score for HS15. Please contact CSDE for other forms and combinations.

Part 5: Test Statistics

5.1. Reliability

Reliability is a statistical index of the consistency of test performance over repeated trials. The simplest model for conveying the concept of reliability is to describe the test re-test method. If a test is administered to a group of examinees and then re-administered to the same examinees a short time later, the correlation of the scores across both test administrations estimates the reliability of the test. To measure reliability using a single administration, the test items are split using various techniques into half-length tests and those scores are then correlated. Cronbach's alpha estimates the lower-bound estimate of an infinite combination of split-halves and therefore is regarded as a very conservative method for assessing test reliability.

Table 8 summarizes reliability estimates for CAPT Mathematics, Science, Reading, and Writing. The reliability coefficients are based on Cronbach's alpha measure of internal consistency. When evaluating these results it is important to remember that reliability is partially a function of test length and thus reliability is likely to be greater for clusters that have more items. Within each content area the reliability estimates across the forms were very similar. Note that there are no Hybrid forms for Mathematics, Reading and Science. Table 9 presents the mean and standard deviation of students' scale scores.

Table 8: CAPT Cronbach's Alpha

Form	Mathematics	Reading	Writing	Science
HS15	0.943	0.833	0.754	0.931
Hybrid1			0.748	
Hybrid2			0.753	

Table 9: CAPT Scale Score Summary Statistics

Subject	Mean	Standard Deviation
Mathematics	252.67	45.48
Reading	243.83	45.69
Writing	258.83	43.35
Science	258.10	49.46

5.2. Classification Consistency and Accuracy

Classification consistency (see Table 10) and accuracy (see Table 11) were measured using the IRT-Class program developed by CASMA (Center for Advanced Studies in Measurement and Assessment) at the University of Iowa. The decision consistency and accuracy was assessed based on the given ability distribution and the difficulty of the items (IRT parameters).

Table 10: Classification Consistency

Content Area	Overall Classification Consistency	Cut 1	Cut 2	Cut 3	Cut 4
Mathematics	0.77610	0.93720	0.94728	0.94628	0.94224
Reading	0.92547	0.94209	0.96290	0.96285	0.96285
Science	0.78978	0.95938	0.95458	0.94002	0.92882
Writing	0.86922	0.93468	0.95624	0.94888	0.96037

Table 11: Classification Accuracy

Content Area	Overall Classification Accuracy	Cut 1	Cut 2	Cut 3	Cut 4
Mathematics	0.83422	0.95484	0.95952	0.96093	0.95871
Reading	0.91873	0.95503	0.96501	0.97465	0.97040
Science	0.84697	0.97145	0.96736	0.95781	0.94964
Writing	0.87886	0.94234	0.96466	0.96498	0.96859

The results of the program show that for the most part, classifications are highly consistent (see Table 10). The consistency ratings at each cut score are generally in the upper 90s. This tends to tail off at the highest cut score (i.e., the upper end of the distributions). The cumulative effect of applying all cut scores simultaneously yields an average consistency of around low to mid 80s. The classification accuracy examinations show (see Table 11), similarly, that the accuracy ratings at each cut score are generally in the upper 90s.

The program also computes the false negative rates for the test, which in effect are an estimate of those students that may have been misclassified in a performance category lower than their true performance category. The results of the false negatives, found in Table 12, indicate that a very small number of students may have been negatively misclassified in this way. Table 13 shows the false positive classification.

Table 12: False Negative Classification

Content Area	Overall False Negative	Cut 1	Cut 2	Cut 3	Cut 4
Mathematics	0.06716	0.01749	0.00998	0.01394	0.02584
Reading	0.04372	0.02985	0.03151	0.01328	0.00601
Science	0.07185	0.01333	0.01263	0.02337	0.02282
Writing	0.08155	0.04316	0.02882	0.02045	0.02157

Table 13: False Positive Classification

Content Area	Overall False Positive	Cut 1	Cut 2	Cut 3	Cut 4
Mathematics	0.09862	0.02766	0.03050	0.02513	0.01545
Reading	0.03755	0.01512	0.00348	0.01208	0.02360
Science	0.08117	0.01522	0.02002	0.01881	0.02754
Writing	0.03959	0.01449	0.00652	0.01457	0.00984

Part 6: CAPT3 Standards

When standards were being established for first generation CAPT, a judgmental standard setting process called Modified Angoff (1971) was employed. Through that process, groups of educators who were familiar with the performance of students at a particular grade level in a particular content area were asked to predict how students who just meet a particular standard (e.g., goal standard) would perform on many different CAPT items. Using the judgment of these groups of educators in consideration with other validity checks, appropriate state goal and remedial standards were recommended by the Department and adopted by the State Board of Education. For the second generation CAPT (CAPT2), the standards were set using a method called Book Mark. In the procedure, all items in the test are arranged from easiest to most difficult. Then a group of educators are asked to mark up to the item at which a student at specific standard could respond to correctly. As in the first generation, the standards set by using the Book Mark method were adopted by the State Board of Education.

The third generation (CAPT3) standards were developed by carrying over the CAPT2 standards as well as department staff working with a CAPT3 Standards Advisory Panel composed of technical experts, district content experts and district research and testing specialists. The CAPT3 standards were set to be as rigorous as the CAPT2 standards. Transferring the standards allowed the Department to maintain the same performance standards for NCLB purposes. The purpose of this section is to summarize the procedures used to accomplish the task of carrying over the standards (see Cizek and Bunch, 2007, for a discussion of standard setting procedures).

In all content areas, the standards define the different academic performance levels. The state goal has been an important benchmark for judging the quality of education in Connecticut for more than a decade. The proficient standard is used for accountability purposes as required by No Child Left Behind (NCLB) to make determinations about Adequate Yearly Progress (AYP) and schools in need of improvement.

To continue to comply with the NCLB accountability requirements, the Connecticut State Department of Education (CSDE) carried over from the CAPT2 to the CAPT3 the following previously adopted achievement standards: Below Basic, Basic, Proficient, Goal and Advanced. The process of carrying over the standards was accomplished with an intergeneration linking study which included the equating of CAPT2 forms and CAPT3 forms. In addition to statistically linking the test generations, historical results from past CAPT2 administrations were taken into consideration as well as input from the CAPT Standards Review Panel composed of a diverse group of Connecticut educators, including curriculum directors, teachers and administrators.

The Standards Review Panel assisted in the identification of acceptable and valid test standards for each content area of CAPT3. The CAPT Standards Review Panel was given an overview of the CAPT3 including the content covered, score weighting, and reporting conventions. Differences between CAPT2 and CAPT3 were also discussed. Copies of the complete CAPT3 test booklets were available for reference. In addition, the procedures for carrying CAPT2 standards over to CAPT3 were presented in detail so that committee members would better understand their role in the process. They reviewed data from several related analyses and discussed implications from both an educational perspective and a technical perspective. They were asked particularly to provide input in the following three areas:

- Review the content of the CAPT, score weighting, and reporting conventions.
- Review results from the inter-generational linking procedure to ensure that standards are reasonable and appropriate across content area; and
- Provide subjective input about the reasonableness and consistency of the standards for all content areas based on their content expertise and historical results from past test administrations.

All procedures were discussed with and approved by the Technical Advisory Committee (TAC) prior to implementation. The TAC is composed of nationally recognized experts in the measurement field. Finally, standards proposed by the standards review panel were presented to the State Board of Education for final approval. Standards were established based on scale scores (100-400) in four content areas: Mathematics, Science, Reading, and Writing.

Table 14 shows the range of scale scores in each performance category.

Table 14: 2008 CAPT Achievement Levels and Scale Score Ranges

Content Area	Scale Score Ranges				
	Below Basic	Basic	Proficient	Goal	Advanced
Mathematics	100 - 190	191 - 220	221 - 259	260 - 289	290 - 400
Science	100 - 189	190 - 214	215 - 264	265 - 294	295 - 400
Reading	100 - 173	174 - 204	205 - 250	251 - 282	283 - 400
Writing	100 - 181	182 - 209	210 - 249	250 - 285	286 - 400

Part 7: Validity

According to the 1999 AERA, APA, NCME *Standards*, "It is helpful to consider the four phases leading from the original statement of purpose(s) to the final product: (a) delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured; (b) development and evaluation of the test specifications; (c) development, field testing, evaluation, and selection of the items and scoring guides and procedures; and (d) the assembly and evaluation of the test for operational use.

In the development and maintenance of CAPT each of these phases is carefully planned and implemented. The following section details the critical psychometric procedures undertaken to ensure a strong validity argument for the use and interpretation of CAPT (Kane, 2006; Messick, 1989).

7.1. Content Validity Survey

In order for the CAPT to serve its intended purposes, it is critical that users of the test results be confident that those results are meaningful. The test must measure those competencies that are critical to the decisions the test scores are informing.

A content validation study was conducted to examine the content validity of the CAPT for its intended applications. For this study, a survey of the strands proposed for the second generation CAPT was sent to approximately 4,000 Connecticut educators, parents, and other citizens. The purpose of the survey was to determine 1) the importance of the proposed Mathematics, Science, Reading Across the Disciplines, and Writing Across the Disciplines strands and 2) whether the strands are taught prior to the end of the 10^{th} grade. The respondents characterized the strands as important educational outcomes to which students would be instructed prior to testing.

7.2. Scoring Quality Assurance Procedures Undertaken during Development

Much of the following discussion applies to procedures undertaken during field testing and test construction phases of development work. Of course quality control is applied during the operational administration, but not with the aim of selecting or removing items.

In order to ensure the validity of inferences made from the CAPT tests there are quality control procedures in place for the scoring of the test. One such quality assurance component is to check the MC answer keys for MC items several times prior to test administration and one final time during the first run of live results. Items yielding low point biserial correlations are checked a final time for miskeying.

For constructed-response (CR) items, CAPT staff and contractor staff work with Connecticut educators to establish score boundaries in a process known as "range finding". The score point examples and training sets so established are carried forward into operational scoring and elaborated with new samples of student responses. Reader training lasts up to several days, and readers must qualify by matching scores to several sets of prescored student responses. Once scoring begins, validity packets are used to maintain reader accuracy. These are packets of student responses with scores pre-assigned by CAPT staff and Connecticut educators. Readers periodically receive these packets, and their responses are compared to the pre-assigned scores. If a reader assigns too many discrepant scores, that reader is retrained or removed from the project. Other QA procedures include a 100% second read for the writing prompts (IW). There is a 20% second read for short answer and extended response items in mathematics and reading comprehension.

7.3. Item Quality Analysis Undertaken During Development

Another part of assessing the quality and validity of inferences made from an instrument is to assess the quality of the items on the test. This quality is typically assessed by examining the classical item statistics as well as the potential for item bias. Item bias could lead to less valid inferences made for certain subgroups.

Item specifications. CAPT employs *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) as a primary source of guidance in the construction, field testing, and documentation of the tests. The introduction to the 1999 *Standards* best describes how those *Standards* are and will be used in the development and evaluation of CAPT tests:

> Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist. (*Standards*, p. 4)

Thus, the terms 'target' and 'goal' are used when referring to various psychometric properties of the tests. For example, while it is a goal of test development for each high school test to have a reliability coefficient of .90 or greater, it is not our intention to scrap a test with a reliability coefficient of .89. Instead, the test results would be published, along with the reliability coefficient and associated standard error of measurement.

Item statistics. Because the CAPT tests are used in making individual decisions about students, they must be very reliable, particularly at cut points (the score points that separate adjacent achievement categories). Target reliability coefficients of .90 (or higher) are therefore set for the important cut points of each test.

Other psychometric properties include item difficulty, item discrimination, and differential item functioning. General statistical targets are provided below:

For Multiple-Choice (MC) Items

Percent correct: greater than or equal to .25
Point biserial correlation with total score: greater than or equal to.20
Mantel-Haenszel: No Category C items (see below)

For Constructed-Response (CR) Items

Difficulty: any level as long as all score points are well represented
Correlation with total score: greater than or equal to.20
Generalized Mantel-Haenszel: No chi-square significant at .05 level of alpha

It should be pointed out that the point biserial correlations for MC items and the correlations for CR items refer to total scores of the field test form with the influence of the item in question removed.

Differential item functioning. The Mantel_Haenszel statistic computes an odds ratio for each item that compares item performance for a reference group and a focal group (for whom bias may be an issue). Specifically, the M-H statistic is a ratio of the probability of success on an item for the reference group to the probability of success on the same item for the reference group. When the ratio is greater than one, the probability of success on the item favors the reference group over the focus group. Note that M-H and other methods for identifying statistical bias are flagging mechanisms that do not necessarily mean that the performance difference is due to unfairness in the item. Instead, the standard procedure is for the bias committee review the items to make a final judgmental determination as to whether or not the item is actually biased.

Since its introduction in the field of epidemiology in 1959, Mantel-Haenszel statistics have been employed by many test developers, and several refinements have been added. Educational Testing Service (ETS) uses the Mantel-Haenszel statistic and calculates a D statistic which permits grouping of test items into three categories (Zieky, 1993). The D statistic is a function of the case-control odds estimator of risk generated by SAS's PROC FREQ. The D statistic is calculated as follows:

1. α = case-control estimate of risk (odds ratio)
2. β = natural log of α
3. $D = -2.35*\beta$

Camilli and Shepard (1994, p. 121) describe three categories of items with respect to D:

A D does not significantly differ from zero using Mantel-Haenszel chi-square, or D's absolute value is less than 1

B D significantly differs from 0 and D has either (a) an absolute value less than 1.5 or (b) an absolute value not significantly different from 1

C D's absolute value is significantly greater than or equal to 1.5

Camilli and Shepard note that Category B items are typically investigated for potential bias, while Category C items are typically removed. Others treat Category C items only as candidates for elimination, pending a reprieve from the committee. In other words, Category C items are considered unusable unless specifically declared usable by the committee. It should be noted that an item that allowed a target group to break out of a pattern of trailing behind the reference group on all other items would tend to fall into Category C. The committee would likely want to keep such an item, in spite of its Mantel-Haenszel status.

DIF occurs when an item shows different results by group (e.g., by race, or sex) that cannot be explained by known differences in the overall achievement levels of the two groups. Overall achievement level is typically taken as scores on an operational test, assuming that the operational test is itself free of bias. While committee members are free to examine all field-tested items, they must review all items with a Category C rating. Unless the committee specifically calls for the inclusion of any such item, that item is removed from the pool.

7.4. Equating Design

A different CAPT form is used each year. In order to ensure that appropriate comparisons can be made from one form of the CAPT to another, test forms must be equivalent to each other. Care must be taken when test items are developed, when items are selected to create forms, when tests are administered, and when tests are scored to keep all conditions as similar as possible for one test form to another. Two important characteristic that must be similar across forms are the content that is measured and the difficulty of the test.

Part 4 of this report details the procedures used to equate and scale the CAPT tests. As mentioned above, three independent groups undertake the analyses and cross-check all analyses and results to ensure accuracy. Connecticut expends great effort and resources to maintain an assessment program that employs high quality psychometric standards and quality assurance.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999*). Standards for educational and psychological testing*. Washington, DC: AERA.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600), Washington, DC: American Council on Research.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Cizek, G. J., & Bunch, M . B. (2007). *Standard setting: A guide to sstablishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 18-64). Westport, CT: American Council on Eduation/Praeger.

Linacre, J. M., & Wright, B. D. (1993, 2006). *A user's guide to BIGSTEPS*. Chicago, IL: MESA Press.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan Publishing Company.

Winsteps. (1991-2006©). Linacre, John M.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 337-364). Hillsdale, NJ: Lawrence Erlbaum Associates.

Appendix A: Item Analysis

Mathematics HS15 Item Analysis

Grid-in Items
PC = Proportion Correct
RPB = Point-Biserial correlation

Open-ended Items
Mean = Mean OE score
Corr = Item-total correlation
0 – 3 = Percent of students at each score point

Item	Type	Rasch	PC/Mean	RPB/Corr
1	OE	-0.3038	1.60	0.44
2	OE	0.7985	1.02	0.67
3	OE	0.5577	1.24	0.53
4	OE	0.0173	1.54	0.68
5	GR	-0.3596	0.57	0.49
6	GR	-0.1993	0.54	0.60
7	GR	0.1938	0.47	0.63
8	GR	0.7846	0.37	0.51
9	GR	-0.1127	0.52	0.63
10	GR	0.6468	0.39	0.51
11	GR	1.2399	0.29	0.41
12	GR	0.9060	0.34	0.61
13	GR	0.6796	0.38	0.48
14	GR	-0.1793	0.54	0.57
15	GR	0.4031	0.43	0.63
16	GR	-0.6775	0.62	0.49
17	OE	-0.4374	1.76	0.57
18	OE	1.0175	0.92	0.52
19	OE	0.1191	1.46	0.55
20	OE	0.1297	1.45	0.77
21	GR	-2.5973	0.87	0.34
22	GR	0.4235	0.43	0.58
23	GR	0.3197	0.45	0.53
24	GR	0.1049	0.49	0.69
25	GR	-0.6546	0.62	0.50
26	GR	0.4021	0.43	0.63
27	GR	-0.1490	0.53	0.62
28	GR	0.1403	0.48	0.61
29	GR	0.8431	0.36	0.44

Item	Type	Rasch	PC/Mean	RPB/Corr
30	GR	-0.3623	0.57	0.50
31	GR	0.1919	0.47	0.68
32	GR	-0.3054	0.56	0.64

Science HS15 Item Analysis

Multiple-choice Items
PC = Proportion Correct
RPB = Point-biserial correlation for keyed answer

Open-ended Items
Mean = Mean OE score
Corr = Item-total correlation
0 – 3 = Percent of students at each score point

Item	Type	Rasch	PC/Mean	RPB/Corr
1	OE	0.6266	1.49	0.57
2	OE	0.4626	1.51	0.59
3	OE	0.3972	1.60	0.58
4	MC	-1.9141	0.89	0.34
5	MC	-0.5035	0.71	0.36
6	MC	-0.4838	0.71	0.50
7	MC	0.6751	0.48	0.33
8	MC	-0.2437	0.66	0.54
9	MC	-0.1538	0.65	0.42
10	MC	1.7534	0.27	0.32
11	MC	0.8046	0.45	0.24
12	MC	0.2424	0.57	0.32
13	MC	-1.3287	0.83	0.39
14	MC	0.6166	0.49	0.41
15	MC	0.6228	0.49	0.42
16	MC	0.6204	0.49	0.39
17	MC	-0.3750	0.69	0.36
18	MC	-1.0361	0.79	0.53
19	MC	-1.3504	0.83	0.49
20	MC	0.2835	0.56	0.48
21	MC	-0.2992	0.67	0.45
22	MC	1.0948	0.40	0.21
23	MC	-0.5294	0.71	0.35
24	MC	-0.0138	0.62	0.34
25	MC	-0.5401	0.72	0.45
26	MC	-0.7326	0.75	0.33
27	MC	0.0292	0.61	0.57
28	MC	0.1783	0.58	0.48
29	MC	0.2054	0.58	0.37
30	MC	0.4534	0.53	0.36
31	MC	0.6955	0.48	0.36

Item	Type	Rasch	PC/Mean	RPB/Corr
32	OE	0.0655	1.86	0.50
33	OE	0.0711	1.87	0.66
34	MC	-1.0001	0.79	0.32
35	MC	-0.8117	0.76	0.37
36	MC	0.6671	0.48	0.31
37	MC	-0.2935	0.67	0.37
38	MC	1.1809	0.38	0.24
39	MC	-0.2405	0.66	0.20
40	MC	-0.6061	0.73	0.43
41	MC	-0.2225	0.66	0.44
42	MC	0.2121	0.58	0.40
43	MC	0.6184	0.49	0.33
44	MC	1.0466	0.41	0.25
45	MC	-1.3098	0.83	0.49
46	MC	0.2036	0.58	0.46
47	MC	0.6590	0.48	0.45
48	MC	1.4649	0.33	0.26
49	MC	-0.1947	0.65	0.36
50	MC	-0.6903	0.74	0.45
51	MC	-0.4411	0.70	0.40
52	MC	0.8632	0.44	0.28
53	MC	-0.4065	0.69	0.38
54	MC	0.2981	0.56	0.25
55	MC	-1.0039	0.79	0.50
56	MC	-0.3388	0.68	0.34
57	MC	0.1338	0.59	0.47
58	MC	0.6887	0.48	0.32
59	MC	0.0640	0.60	0.28
60	MC	0.1826	0.58	0.43
61	MC	-0.1194	0.64	0.37
62	MC	0.3049	0.56	0.46
63	MC	-0.1282	0.64	0.50
64	MC	-0.6449	0.73	0.50
65	MC	-0.2375	0.66	0.42

Reading for Information HS15 Item Analysis

Multiple-choice Items
PC = Proportion Correct
RPB = Point-biserial correlation for keyed answer

Open-ended Items
Mean = Mean OE score
Corr = Item-total correlation
0 – 2 = Percent of students at each score point

Item	Type	Rasch	PC/Mean	RPB/Corr
1	MC	-1.9486	0.86	0.31
2	MC	-2.5319	0.91	0.37
3	MC	-0.4514	0.65	0.34
4	MC	-1.5318	0.81	0.34
5	OE	1.0809	0.71	0.43
6	OE	0.5465	0.94	0.53
7	MC	-0.7432	0.70	0.38
8	MC	-0.0893	0.58	0.33
9	MC	-1.2901	0.78	0.35
10	MC	-0.1053	0.58	0.43
11	OE	0.4723	0.95	0.51
12	OE	1.6293	0.62	0.51
13	MC	-1.4871	0.81	0.43
14	MC	0.7592	0.42	0.33
15	MC	-1.0903	0.75	0.36
16	MC	-0.6005	0.67	0.45
17	OE	1.5988	0.65	0.48
18	OE	2.0515	0.41	0.39

Editing and Revising HS15 Item Analysis

Multiple-choice Items
PC = Proportion Correct
RPB = Point-biserial correlation for keyed answer

Item	Type	Rasch	PC	RPB
1	MC	1.0055	0.56	0.19
2	MC	-1.0504	0.88	0.23
3	MC	1.9577	0.37	0.23
4	MC	-0.2414	0.78	0.35
5	MC	0.9811	0.57	0.23
6	MC	-1.3144	0.90	0.36
7	MC	-1.7100	0.93	0.27
8	MC	1.4166	0.48	0.24
9	MC	1.3188	0.50	0.32
10	MC	0.4547	0.67	0.37
11	MC	-0.9919	0.87	0.35
12	MC	1.4881	0.46	0.20
13	MC	-1.5200	0.92	0.38
14	MC	-0.2586	0.79	0.29
15	MC	-1.3221	0.90	0.32
16	MC	0.8309	0.60	0.23
17	MC	0.1986	0.71	0.31
18	MC	1.6623	0.43	0.17

Response to Literature and Interdisciplinary Writing HS15 Item Analysis

Extended Response
Mean = Mean EX score
Corr = Item-total correlation
2 – 12 = Percent of students at each point

	Type	Rasch	Mean	Corr	2	3	4	5	6	7	8	9	10	11	12
RL	EX	0.4173	7.08	0.62	0.01	0.01	0.06	0.07	0.21	0.18	0.25	0.12	0.06	0.01	0.00
IW1	EX	1.2648	7.50	0.71	0.02	0.02	0.05	0.05	0.13	0.13	0.31	0.15	0.11	0.02	0.01
IW2	EX	1.4465	7.21	0.69	0.02	0.02	0.06	0.07	0.15	0.16	0.30	0.13	0.08	0.01	0.00

Appendix B: Raw, Theta, and Scale Scores

Raw, Theta, and Scale Scores for Mathematics HS15

Raw Score	Theta	Scale Score
0	-5.3326	110
1	-4.0562	146
2	-3.2701	168
3	-2.7842	182
4	-2.4267	192
5	-2.1425	200
6	-1.9060	206
7	-1.7031	212
8	-1.5252	217
9	-1.3664	221
10	-1.2230	225
11	-1.0919	229
12	-0.9711	232
13	-0.8587	236
14	-0.7536	239
15	-0.6546	241
16	-0.5607	244
17	-0.4712	246
18	-0.3852	249
19	-0.3022	251
20	-0.2215	253
21	-0.1426	256
22	-0.0651	258
23	0.0114	260
24	0.0876	262
25	0.1637	264
26	0.2401	266
27	0.3172	269
28	0.3953	271
29	0.4748	273
30	0.5561	275
31	0.6397	278
32	0.7258	280
33	0.8152	282
34	0.9085	285
35	1.0062	288
36	1.1095	291
37	1.2193	294
38	1.3372	297
39	1.4650	301
40	1.6053	305
41	1.7615	309
42	1.9387	314
43	2.1446	320
44	2.3923	327
45	2.7055	335
46	3.1378	348
47	3.8593	368
48	5.0843	400

Raw, Theta, and Scale Scores for Science HS15

Raw Score	Theta	Scale Score
0	-5.6818	100
1	-4.4612	100
2	-3.7444	100
3	-3.3154	100
4	-3.0043	100
5	-2.7579	100
6	-2.5523	106
7	-2.3749	115
8	-2.2182	123
9	-2.0772	129
10	-1.9486	136
11	-1.8301	141
12	-1.7199	147
13	-1.6166	152
14	-1.5191	157
15	-1.4267	161
16	-1.3387	165
17	-1.2546	169
18	-1.1737	173
19	-1.0959	177
20	-1.0207	181
21	-0.9479	184
22	-0.8772	188
23	-0.8084	191
24	-0.7413	194
25	-0.6757	198
26	-0.6116	201
27	-0.5486	204
28	-0.4868	207
29	-0.4260	210
30	-0.3660	213
31	-0.3069	215
32	-0.2483	218
33	-0.1904	221
34	-0.1330	224
35	-0.0760	227
36	-0.0193	229
37	0.0372	232
38	0.0935	235
39	0.1497	238
40	0.2060	240
41	0.2624	243
42	0.3189	246
43	0.3758	249
44	0.4330	251
45	0.4907	254
46	0.5490	257
47	0.6079	260
48	0.6676	263
49	0.7283	266
50	0.7900	269
51	0.8529	272
52	0.9171	275
53	0.9829	278
54	1.0504	281
55	1.1198	285
56	1.1914	288
57	1.2656	292
58	1.3425	296
59	1.4227	300
60	1.5065	304
61	1.5945	308
62	1.6875	312
63	1.7862	317
64	1.8918	322
65	2.0054	328
66	2.1290	334
67	2.2649	341
68	2.4166	348
69	2.5887	356
70	2.7891	366
71	3.0302	378
72	3.3360	393
73	3.7596	400
74	4.4711	400
75	5.6880	400

Raw, Theta, and Scale Scores for Reading HS15

Raw Score	Theta	Scale Score
0	-5.3511	100
1	-4.1925	100
2	-3.5403	100
3	-3.1548	110
4	-2.8709	121
5	-2.6384	129
6	-2.4359	137
7	-2.2519	143
8	-2.0802	150
9	-1.9169	156
10	-1.7596	161
11	-1.6064	167
12	-1.4562	172
13	-1.3081	178
14	-1.1613	183
15	-1.0152	189
16	-0.8695	194
17	-0.7239	199
18	-0.5783	205
19	-0.4327	210
20	-0.2871	215
21	-0.1415	221
22	0.0040	226
23	0.1498	231
24	0.2962	237
25	0.4437	242
26	0.5928	248
27	0.7442	253
28	0.8985	259
29	1.0561	265
30	1.2176	270
31	1.3831	277
32	1.5528	283
33	1.7269	289
34	1.9054	296
35	2.0889	302
36	2.2781	309
37	2.4740	317
38	2.6780	324
39	2.8915	332
40	3.1161	340
41	3.3534	349
42	3.6056	358
43	3.8771	368
44	4.1773	379
45	4.5261	392
46	4.9730	400
47	5.6816	400
48	6.8739	400

Raw, Theta, and Scale Scores for Writing HS15

Raw Score	Theta	Scale Score
0	-5.0321	100
1	-3.7957	100
2	-3.0656	100
3	-2.6328	112
4	-2.3253	125
5	-2.0876	134
6	-1.8939	142
7	-1.7302	148
8	-1.5879	154
9	-1.4612	159
10	-1.3462	164
11	-1.2404	168
12	-1.1415	172
13	-1.0481	176
14	-0.9590	180
15	-0.8730	183
16	-0.7897	186
17	-0.7082	190
18	-0.6280	193
19	-0.5486	196
20	-0.4697	199
21	-0.3909	202
22	-0.3119	206
23	-0.2323	209
24	-0.1519	212
25	-0.0704	215
26	0.0125	219
27	0.0971	222
28	0.1837	225
29	0.2725	229
30	0.3640	233
31	0.4586	237
32	0.5566	240
33	0.6585	245
34	0.7647	249
35	0.8756	253
36	0.9916	258
37	1.1127	263
38	1.2392	268
39	1.3708	273
40	1.5074	279
41	1.6484	284
42	1.7934	290
43	1.9422	296
44	2.0948	302
45	2.2512	309
46	2.4121	315
47	2.5784	322
48	2.7508	329
49	2.9300	336
50	3.1164	343
51	3.3097	351
52	3.5097	359
53	3.7165	368
54	3.9317	376
55	4.1595	385
56	4.4094	395
57	4.7002	400
58	5.0766	400
59	5.6880	400
60	6.7697	400